ANALYSIS OF MAIN OTU PICKING METHODS BY BIPARTITE GRAPHS

Marcela Šafárová

Master Degree Programme (2), FEEC BUT

E-mail: xsafar15@stud.feec.vutbr.cz

Supervised by: Karel Sedlář

E-mail: sedlar@feec.vutbr.cz

Abstract: Studies of ecosystems containing millions of microbial organisms, so called microbiomes, using DNA sequencing became very popular. The main step involves fast clustering of sequences belonging to the same operational taxonomic units (OTU) that represent the same or taxonomically related organisms. This step, known as OTU picking, can be performed by several techniques and significantly affects the final interpretation of the data. In this paper, a novel R/Biconductor package for microbiome data presentation using bipartite graphs is introduced while its benefits are demonstrated during comparison of 3 OTU picking techniques.

Keywords: bipartite graphs, microbiome diversity, OTU picking

1 ÚVOD

S rostoucími technologickými možnostmi na poli sekvenování roste také zájem o detailnější zkoumání mikrobiomů (souborů mikroorganismů nacházejících se v daném prostředí) a ukazuje se, že jejich rozdílná diverzita může zásadním způsobem ovlivnit vlastnosti ekosystému. Tendence porovnávat jednotlivé mikrobiomy z hlediska diverzity však naráží na problém nedostatečných nástrojů pro jejich analýzu. Tato práce ukazuje, jak je možné pro analýzy využít bipartitních grafů a zaměřuje se na porovnání hlavních metod zpracování mikrobiálních dat.

2 BIPARTITNÍ GRAF

Bipartitní grafy jsou charakterizovány množinou vrcholů *V*, která může být rozdělena na dvě disjunktní množiny V_1 a V_2, takovým způsobem, že žádné dva vrcholy ze stejné množiny nejsou spojeny hranou *e* náležící do množiny hran *E* bipartitního grafu:

$$V = V_1 \cup V_2, V_1 \cap V_2 = \emptyset \quad (1)$$

$$\forall e = \{u, v\}, e \in E: u \in V_1 \wedge v \in V_2 \quad (2)$$

Disjunktní množiny V_1 a V_2 se také nazývají partity. Mikrobiální data mohou být snadno rozdělena do dvou partit, a tak efektivně vizualizována pomocí bipartitních grafů. První partita je obvykle tvořena analyzovanými vzorky, druhá partita pak pozorovanými mikroorganismy **Chyba! Nenalezen zdroj odkazů.**.

3 IMPLEMENTACE METODY

V programovacím prostředí R byl vytvořen balíček bipartiteOTU, umožňující zpracování a čištění kvantitativních mikrobiálních dat a následnou analýzu pomocí bipartitních grafů. Balíček nabízí také možnost modifikace grafů pomocí váhování hran a uzlů, detekce komunit a barvení vrcholů a hran. Bipartitní grafy mohou být vykresleny přímo v prostředí R nebo uloženy v Graph Modeling Language formátu a importovány do celé řady vizualizačních programů. R balíček je dostupný na www.github.com/safma/bipartiteOTU a jednoduše instalovatelný, včetně zabudované nápovědy, v prostředí R pomocí devtools příkazu install_github("safma/bipartiteOTU").

4 ANALÝZA PROCESU OTU PICKING BIPARTITNÍMI GRAFY

Velmi důležitým krokem analýzy jakýchkoliv mikrobiálních dat je proces porovnávání čtení ze sekvenátoru a jejich shlukování na základě podobností do tzv. OTU (operačních taxonomických jednotek), které obvykle reprezentují jednotlivé mikroorganismy. Postup se nazývá OTU picking a existuje ve třech hlavních podobách:

- *De nov*o OTU picking – Výpočetně nejnáročnější metoda, čtení jsou porovnávána mezi sebou a na základě podobností přiřazována do shluků.
- Closed reference OTU picking – Výpočetně nejméně náročná metoda, čtení jsou porovnávána s referenční databází a přiřazována na základě podobnosti k sekvencím v databázi. Čtení, která nesplňují kritéria pro přiřazení k referenčním sekvencím, jsou z dalšího zpracování vyřazena.
- Open reference OTU picking – čtení jsou porovnávána s referenční databází, avšak sekvence, které nebyly spárovány s referenčními sekvencemi, nejsou vyřazeny, ale shlukovány *de novo* mezi sebou.

OTU picking metody byly srovnány pomocí testovacího datasetu mikrobiomů vyizolovaných z vody využívané k přepravě ryb. Dataset byl publikován v [2]. Jednotlivá čtení byla získána pyrosekvenováním V3/V4 variabilních regionů 16S rRNA genu. Tato data byla prostřednictvím platformy QIIME [3] a výpočetních serverů MetaCentra zpracována zvlášť closed reference, open reference metodou a *de novo* metodou, u které byly výsledné shluky dodatečně porovnány s databází, aby bylo možno získaná data srovnat mezi sebou. Ve všech případech byla jako referenční databáze volena Greengenes 13_8 a shlukovací algoritmus uclust. Jako práh podobnosti pro přiřazení do shluků byla volena hranice 97 %. Výsledky byly pro vizuální analýzu zpracovány do podoby bipartitních grafů.

5 VÝSLEDKY

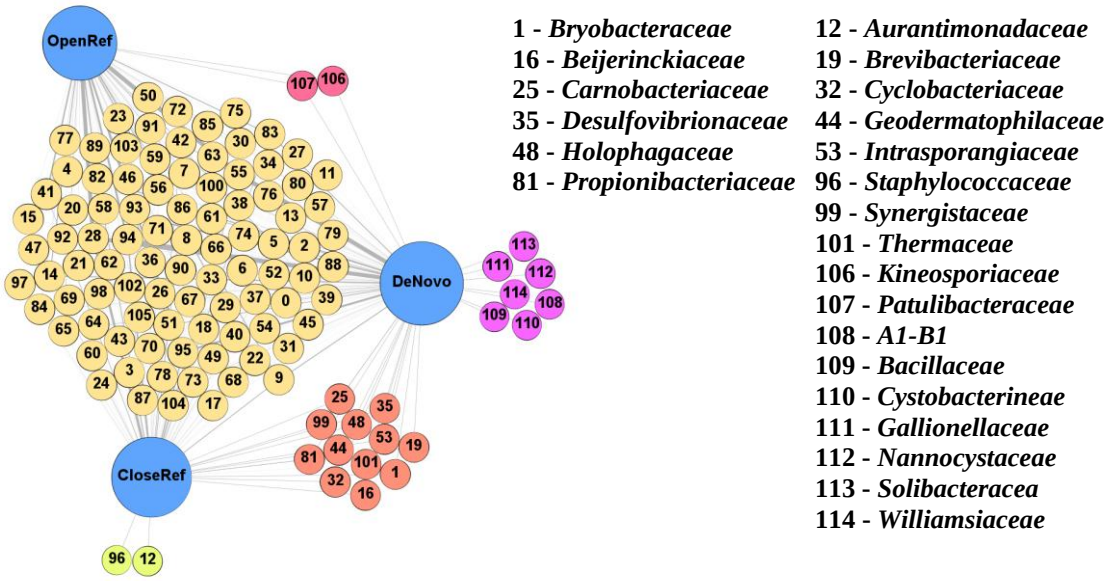
Napříč OTU picking metodami byly zaznamenány rozdíly v celkové detekované mikrobiální diverzitě vzorků. Přičemž odlišnosti ve výsledcích najdeme již na taxonomické úrovni kmene. Zatímco pomocí open reference OTU picking bylo nalezeno 16 mikrobiálních kmenů, přístup closed reference a *de novo* picking jich detekoval 19. Jak je patrné v **Chyba! Nenalezen zdroj odkazů.**, rozdíly ve výsledné diverzitě narůstaly s konkrétnějším taxonomickým zařazením. Obrázek 1 ukazuje odlišnosti v dosažené diverzitě na úrovni čeledi. OTU jsou rozděleny do pěti barevně odlišených komunit na základě příslušnosti k jednotlivým OTU picking metodám. Získaná nižší diverzita open reference metody je zapříčiněna odlišnými defaultními parametry. Open reference totiž využívá prahování a odstraňuje OTU tvořená jediným čtením. Na úrovni taxonomických čeledí se toto nastavení projevilo například potlačením čeledi *Staphylococcaceae* s počtem 11 čtení nalezených closed reference metodou.

OTU picking metoda	shluky	Nová OTU	Celkový počet čtení	Nalezené taxonomické kategorie						
				říše	kmen	třída	řád	čeleď	rod	druh
De novo	4149	921	56661	1	19	38	63	113	185	51
Open reference	2596	276	53740	1	16	31	56	94	134	45
Closed reference	1410	0	30345	1	19	37	62	106	194	89

Tabulka 1: Srovnání výsledků získaných *de novo,* open reference a closed reference picking OTU metodou

Metody se také lišily v počtu nově detekovaných OTU. Z podstaty closed reference přístupu vyplývá, že nemůže detekovat žádné nové OTU a rovněž celkový počet zpracovaných čtení je nižší,

neboť čtení, která nemohou být spárována s referencí, jsou z datasetu vyřazena. Naproti tomu open reference přístup vytvořil 276 nových OTU a *de novo* metoda dokonce 921.



Obrázek 1: Taxonomické čeledi nalezené closed reference, open reference a *de novo* OTU picking metodami.

6 ZÁVĚR

Analýza OTU picking procesu, nezbytného kroku mikrobiálních studií, ukázala, že existují značné rozdíly ve výstupních datech získaných *de novo*, closed reference a open reference metodami. Rozdíly se projevují nejen v celkovém počtu zpracovaných čtení a nalezených taxonů, ale také v počtu nových OTU, ke kterým nemohlo být přiřazené žádné taxonomické určení. Tuto skutečnost je třeba zohlednit zejména při srovnávání výsledků z více studií, neboť nezohlednění OTU picking metody může způsobit milnou interpretaci výsledků.

K analýze bylo využito navrženého R balíčku pro tvorbu bipartitních grafů z kvantitativních mikrobiálních dat. Vizualizace pomocí bipartitních grafů je přehledná a výpočetně nenáročná metoda, která umožňuje nahlédnout do složení jednotlivých mikrobiomů a tím je lépe porovnat mezi sebou.

PODĚKOVÁNÍ

Poděkování patří MetaCentru VO za poskytnutí výpočetních a úložných kapacit, díky nimž mohl být tento projekt zrealizován.

REFERENCE

[1] SEDLAR, Karel, et al. Bipartite graphs for metagenomic data analysis and visualization. In: Bioinformatics *and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015. p. 1123-1128.

[2] GERZOVA, Lenka, et al. Characterization of microbiota composition and presence of selected antibiotic resistance genes in carriage water of ornamental fish. PloS one, 2014, 9.8: e103865.

[3] CAPORASO, J. Gregory, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods, 2010, 7.5: 335-336.